Clustering & PCA Assignment

Overall approach of the analysis by: Sachin Katiyar

Problem Statement

- To categorise the countries based on some socio-economic and health factors that determine the overall development of the country and suggest the countries to the CEO which are in the direst need of aid.

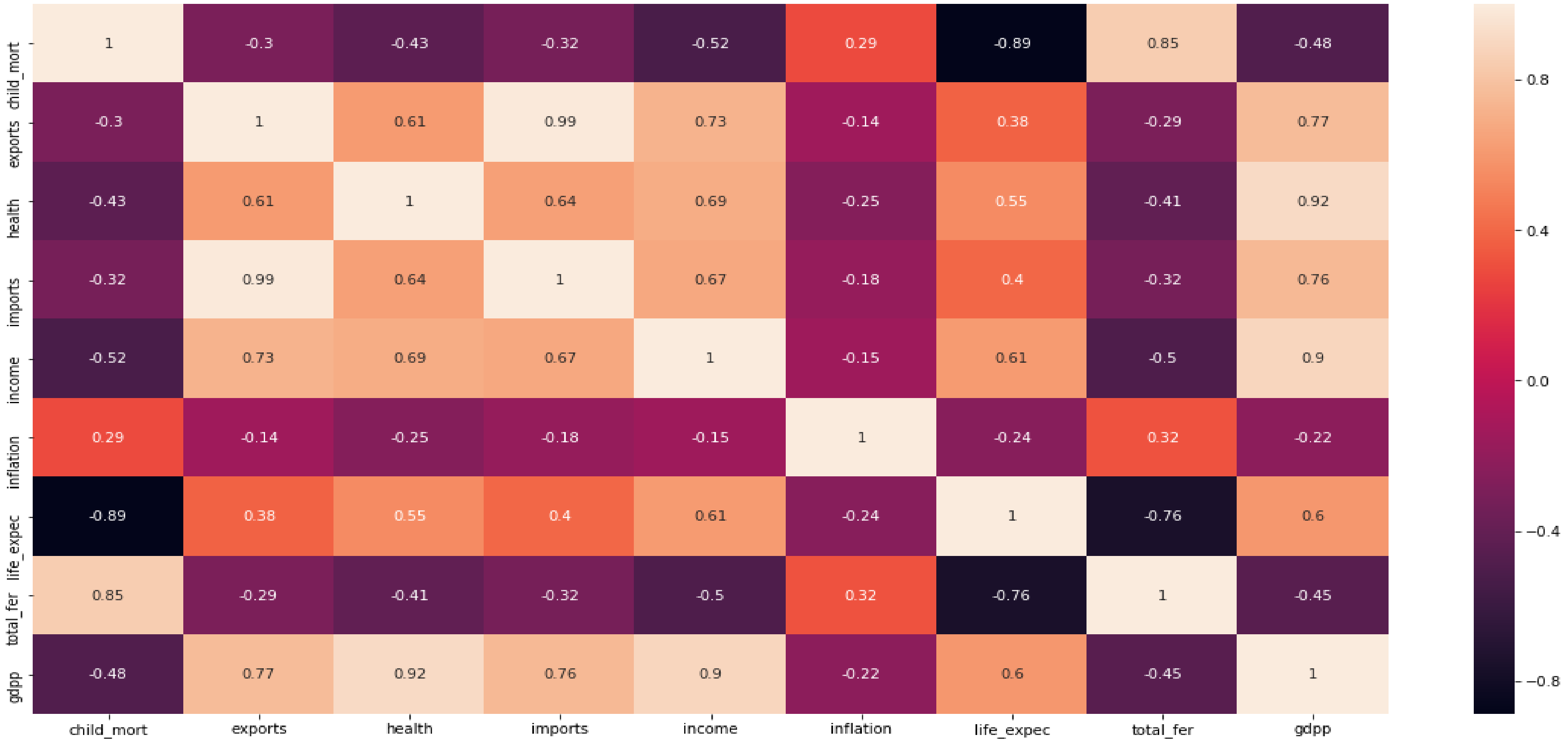
Analysis Approach

Below are the approach taken for analysis the data and finding the required cluster of countries that are in the direst need of aid:

1. Loaded the provided data in our dataframe.
2. Described the provided data for getting data insight.
   - Columns in dataframe are country, child_mort, exports, imports, health, income, inflation, life_expec, total_fer and gdpp
   - No null field present in data
   - Since health, imports and exports are given as percentage of gdpp in provided data therefore converting them to actual values for proper analysis.
   - Then plotted boxplot for all variables to check outliers. A lot of outlier exists but we can't remove them as they are the ones that will lead us to the result. For eg: outliers in life_expec are for the countries which are in need of aid.
   - Also our dataset is very small, removing these outliers will leave us with very less data to analyse on.

Contd…

3. Done Bivariate analysis on numerical columns to check for correlation between variables.



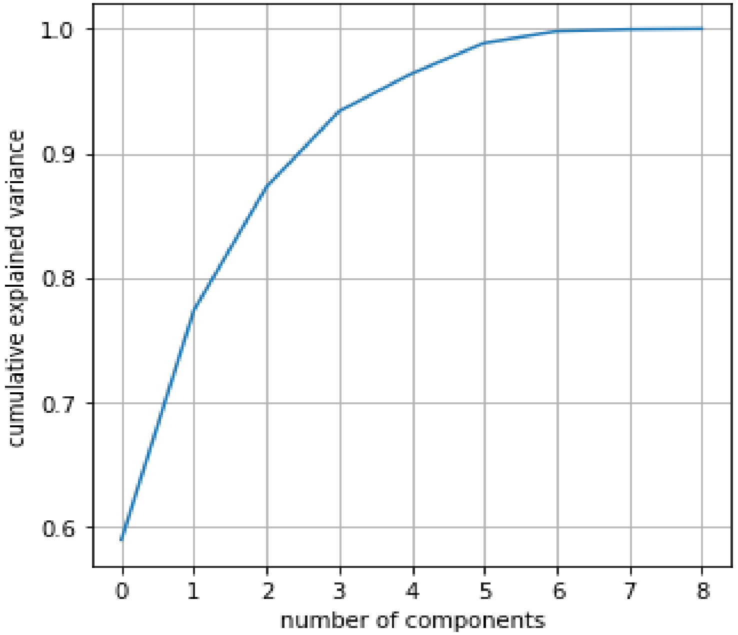
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	0.73	0.69	0.67	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

Contd…

- From the correlation matrix obtained, correlation can be observed between below points:
    - income is highly positively correlated with gdp which means high income of individuals will help in increasing gdp of the country.
    - total-fertility is also highly positively correlated with child mortality which shows that if a women gives many births then death of children below 5 years will be much more.
    - child mortality is highly negatively correlated with life expectancy which shows that if death of children below 5 years is much more then average no of years a child live will be much less.
    - total fertility is highly negatively correlated with life expectancy which shows that if a women gives many births then average no of years a child live will be much less.

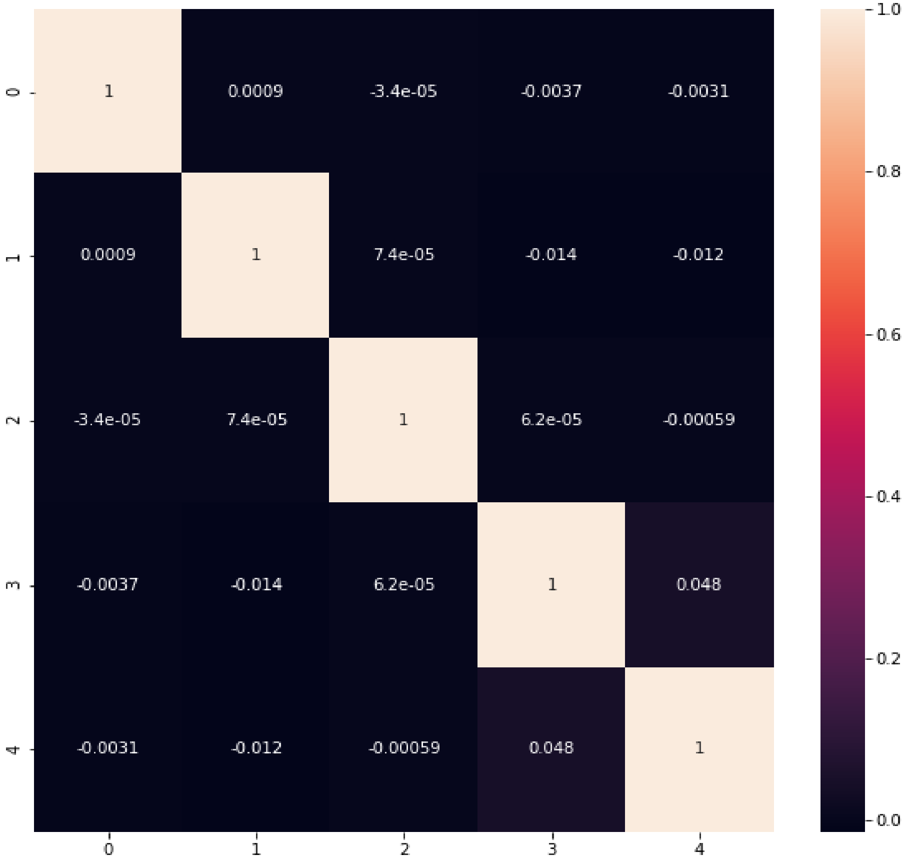
Contd…

4. Next we have done PCA after scaling the data.
    - From the graph obtained, we can see that 96% of variance is explained by 5 components, therefore proceeding with 5 components. PC1 explains maximum variance followed by PC2.



Contd…

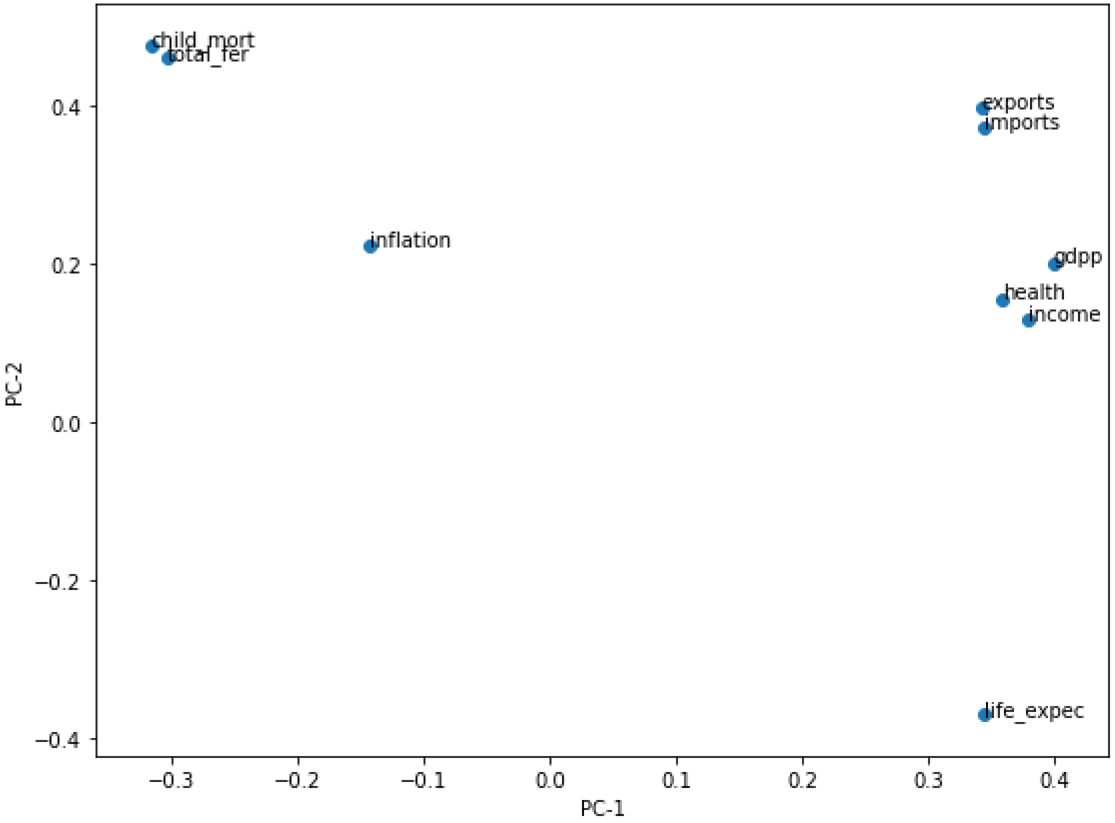
5. Now we checked for correlation between all the 5 components. There should be no correlation between these components and after plotting heat map we observed the same.
   - Min Correlation: -0.0137
   - Max Correlation: 0.0475



	0	1	2	3	4
0	1	0.0009	-3.4e-05	-0.0037	-0.0031
1	0.0009	1	7.4e-05	-0.014	-0.012
2	-3.4e-05	7.4e-05	1	6.2e-05	-0.00059
3	-0.0037	-0.014	6.2e-05	1	0.048
4	-0.0031	-0.012	-0.00059	0.048	1

Contd…

6. Next we have plotted variables against Principal Component 1 and Principal Component 2

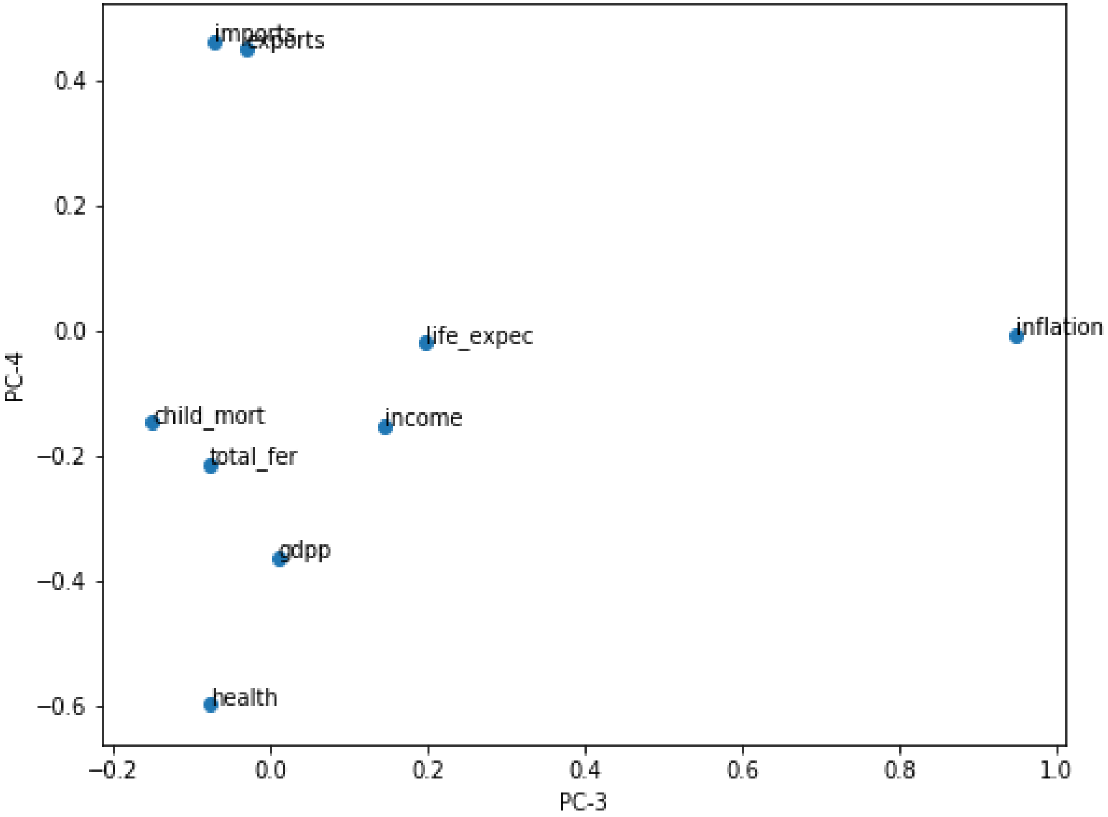


Contd…

- From the above scatter plot we observed below points:
  - PC_1 is in the direction of life_expec, income, gdpp and health. This means that PC_1 value is more for these variables. Also we know that more the value of life_expec, income, gdpp and health, better will be the nation in terms of development. Therefore PC_1 is a good factor for development.
  - PC_2 is in the direction of child_mort and total_fer. This means that PC_2 value is more for these variables. Also we know that less the value of child_mort and total_fer, better will be the nation in terms of development. Therefore PC_2 is a bad factor for development.

Contd…

7. Next we have plotted variables against Principal Component 3 and Principal Component 4

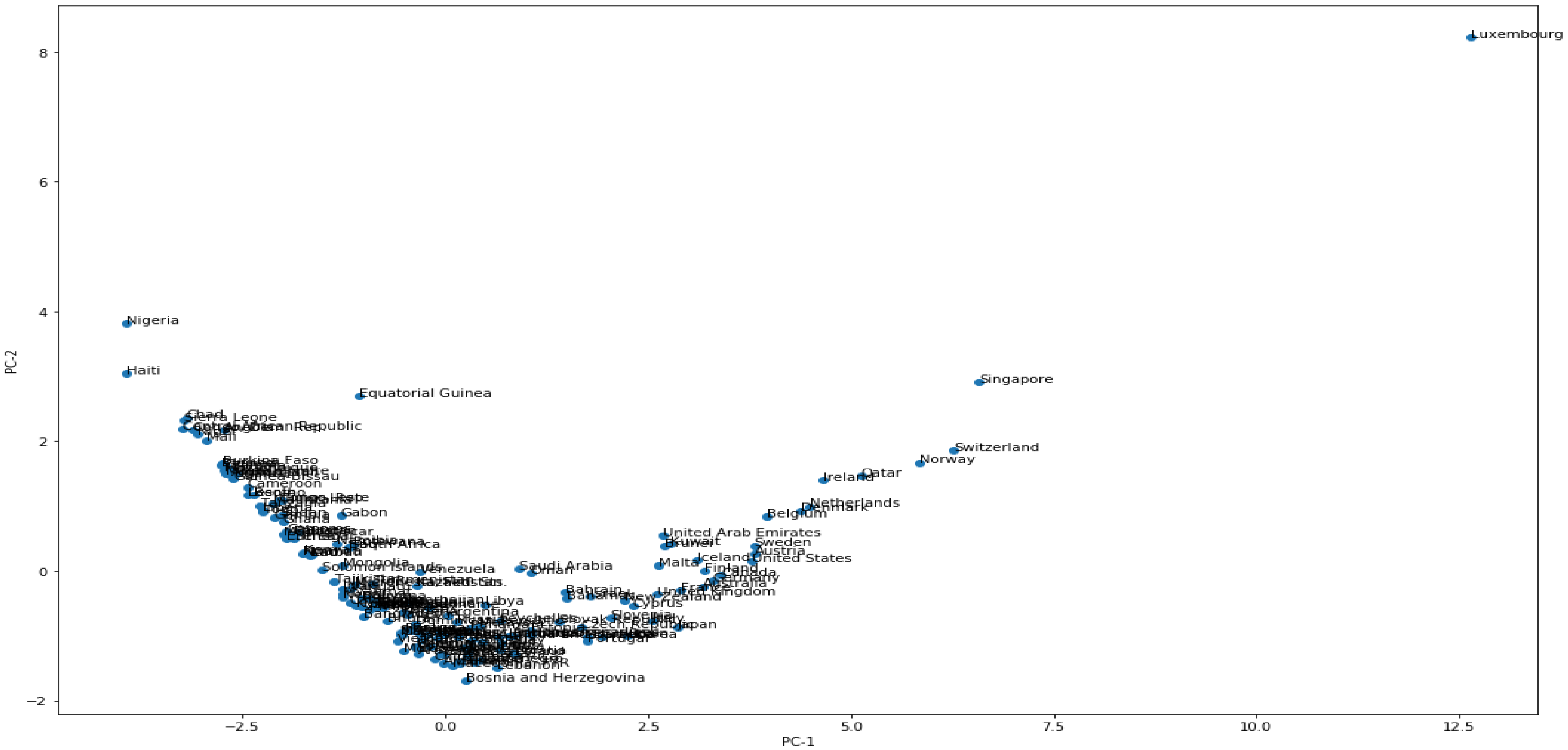


Contd…

- From the above scatter plot we observed below points:
    - PC_3 is in the direction of inflation. This means that PC_3 value is more for this variable. Also we know that less the value inflation, better will be the nation in terms of development. Therefore PC_3 is a bad factor for development.
    - PC_4 is in the direction of exports and imports. This means that PC_4 value is more for these variables. Also we know that more the value exports and imports, better will be the nation in terms of development. Therefore PC_4 is a good factor for development.

Contd…

Plot for Countries against PC1 and PC2



Contd…

8. From above graph we can say that countries which are on top leftmost corner are the ones which are in need of aid as for them PC_2 value is high and PC_1 value is low. If PC_2 value is high then it means that "child_mort and total_fer" is more and if PC_1 value is low then it means that "life_expec, income, gdpp and health" having low values.
   - Countries that are in top leftmost corner are:
     - Haiti
     - Chad
     - Equatorial Guinea
     - Nigeria
     - Sierra Leone

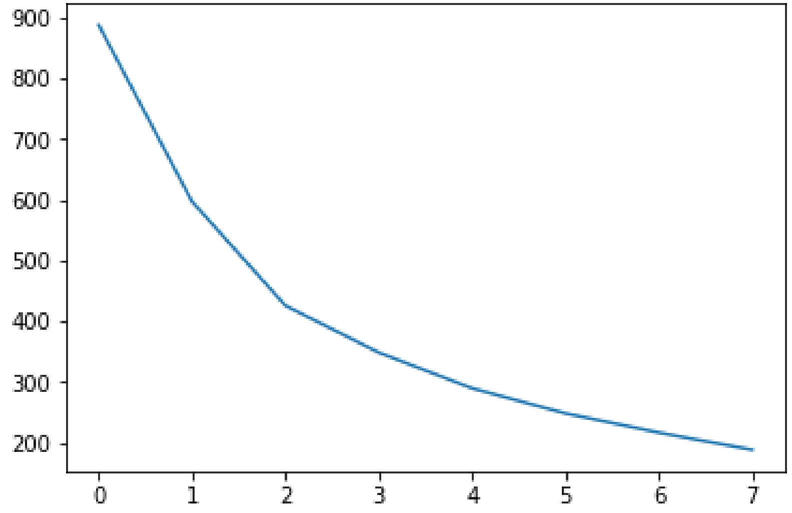
Contd…

9. Next we have done Hopkins Statistics to check whether data provided is good for clustering or not. Since value received from Hopkins Statistics is more than 0.90 therefore we can conclude that data is good for clustering.

Contd…

10. K-Means Clustering:

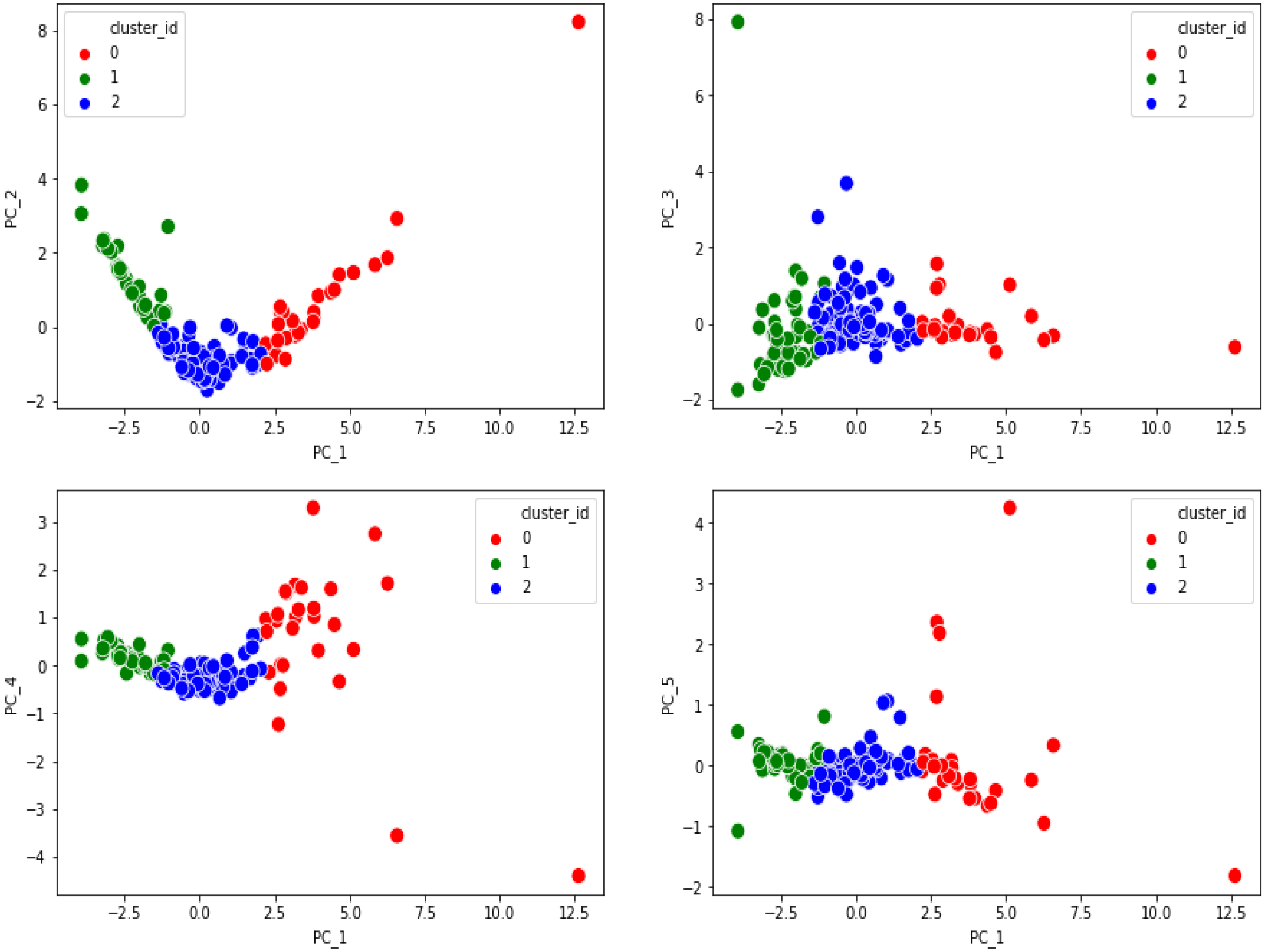
For finding optimal number of clusters, we draw elbow curve and calculated Silhouette Score. Based on Silhouette score and Elbow curve, K = 3 seems good as there is much bend in elbow curve so we will proceed with K = 3 in K-Means Thinking about the business perspective, three groups like Developed Countries, Developing Countries and Under-Developed Countries are the perfect division boundaries between countries. Hence K = 3 seems a reasonable choice from both Statistical and Business view.



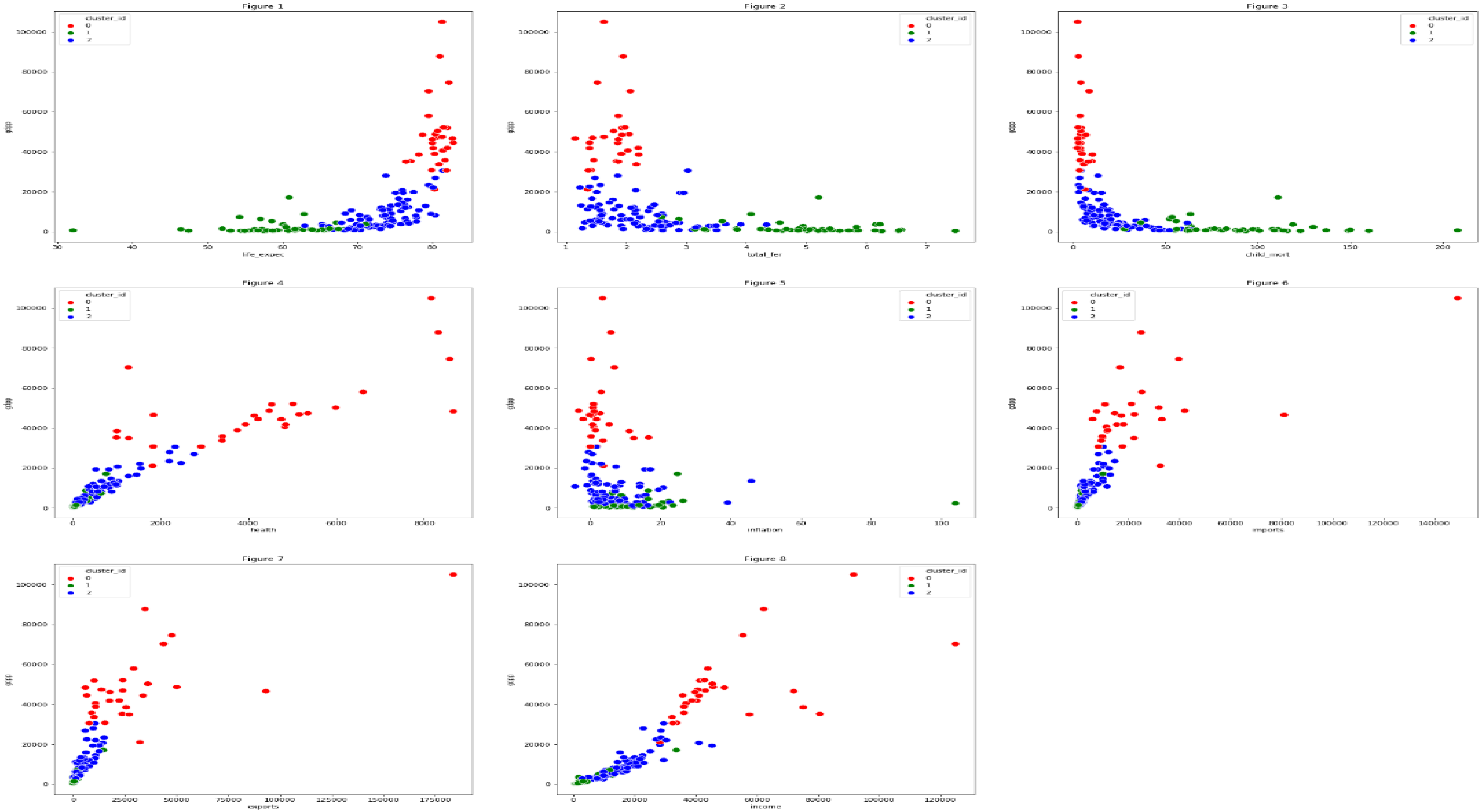
Contd…

11. Next we plot scatterplot between PC1-PC2, PC1-PC3, PC1-PC4 and PC1-PC5, taking PC1 with each as PC1 explains most of the variance.

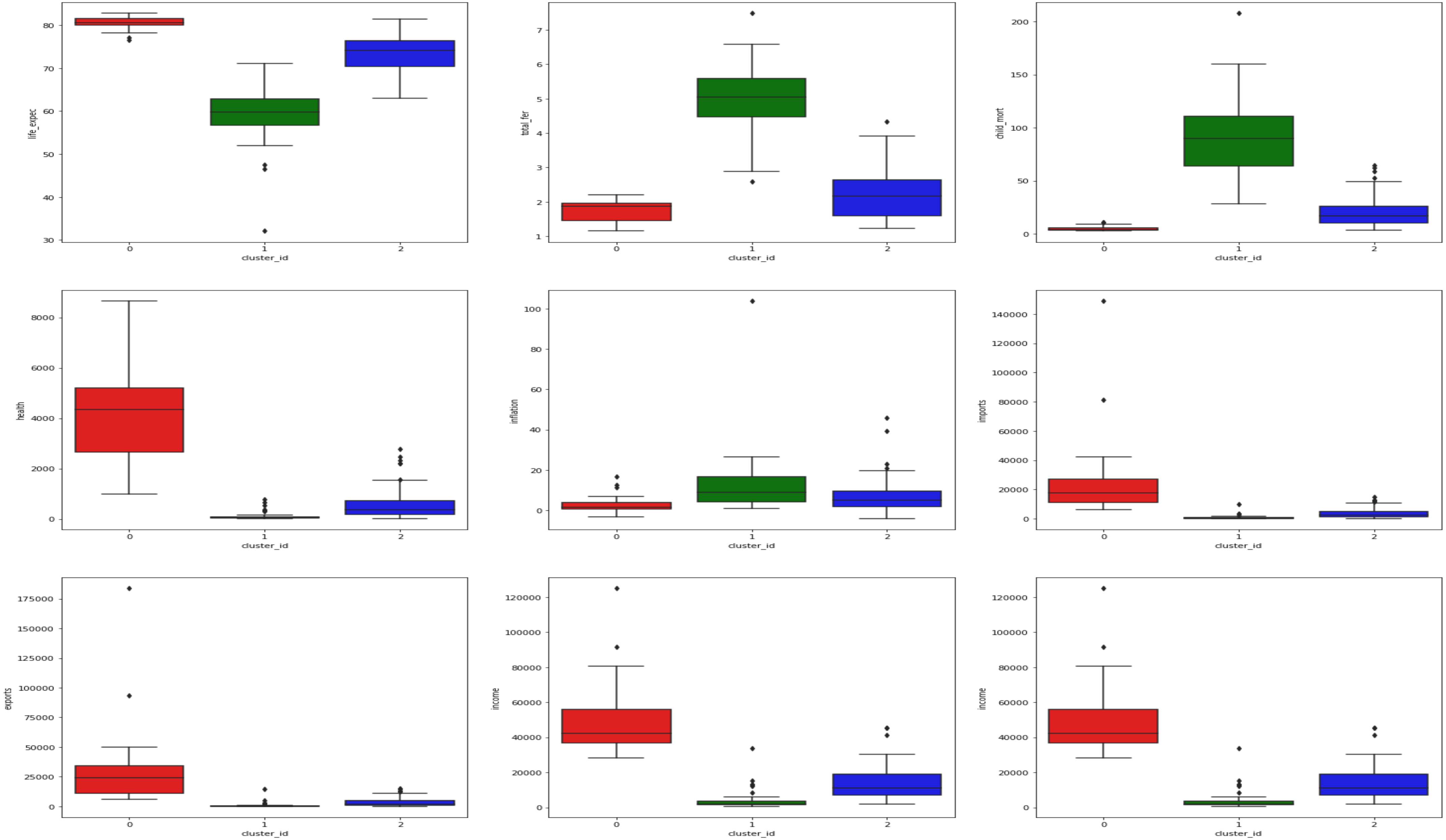
12. It seems that cluster 1 countries are the under-developed countries.



13. Next, we have plot scatterplot between different variables taking hue as cluster id



14. Plotting boxplots on all variables to get better idea of clusters and to give final analysis on the same

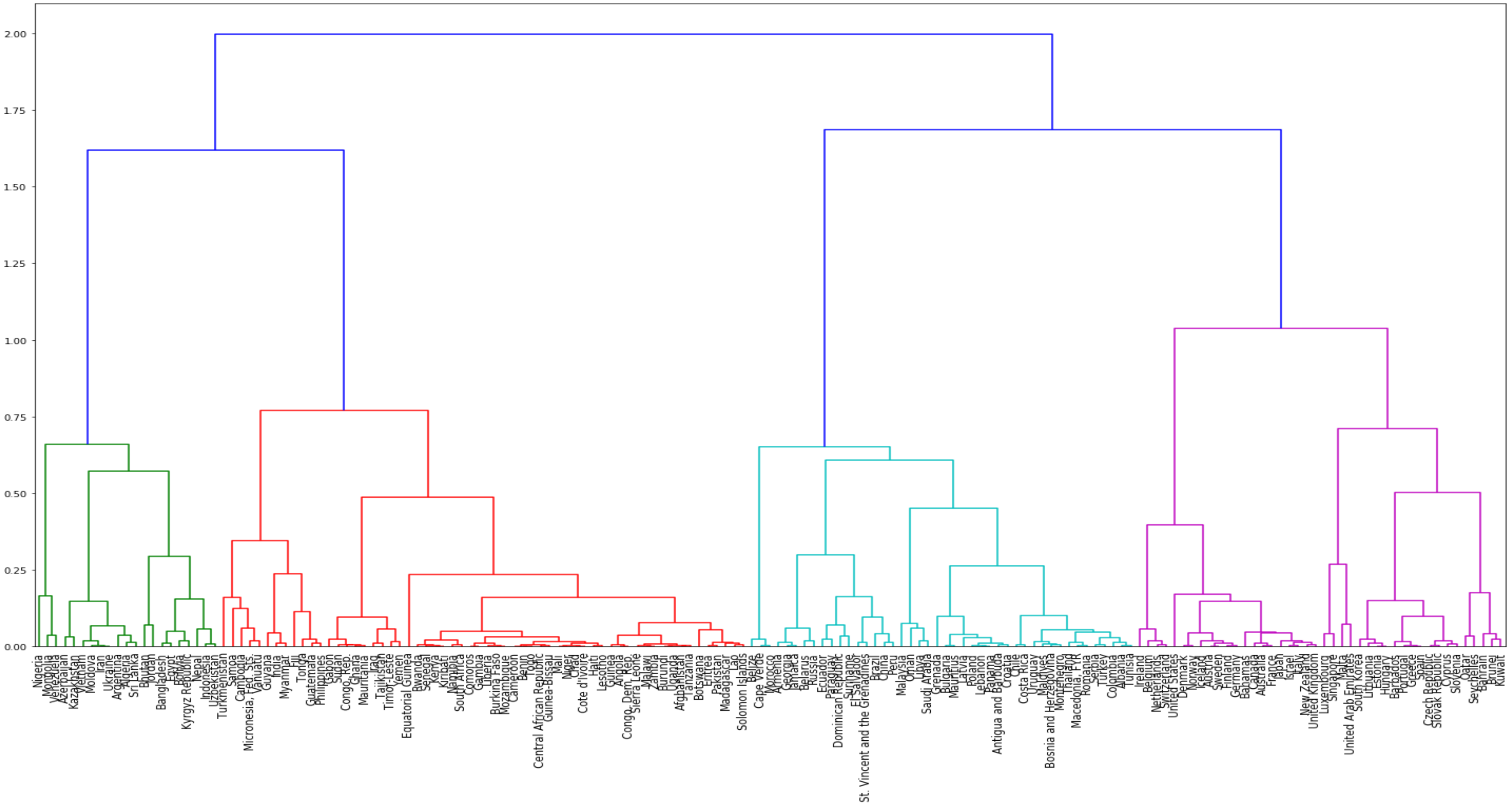


Contd…

- From the above scatter plot and box plots, it has been clear that:
    - Cluster 0 represents developed countries (Blue dot in Scatter plot and Red Box-plot)
    - Cluster 1 represents under-developed countries (Green dot in Scatter plot and Green Box-plot)
    - Cluster 2 represents developing countries (Red dot in Scatter plot and Blue Box-plot)
- Cluster 1 has low life_expec, high_fertility, high child mortality, low health, low imports/exports, low income and low GDP. These all factors shows that Cluster 1 belongs to under-developed countries. After filtering out top 10 countries which are very much in need of aid are:
    - Benin, Burkina Faso, Central African Republic, Congo Dem. Rep., Guinea, Guinea-Bissau, Haiti, Mozambique, Niger and Sierra Leone

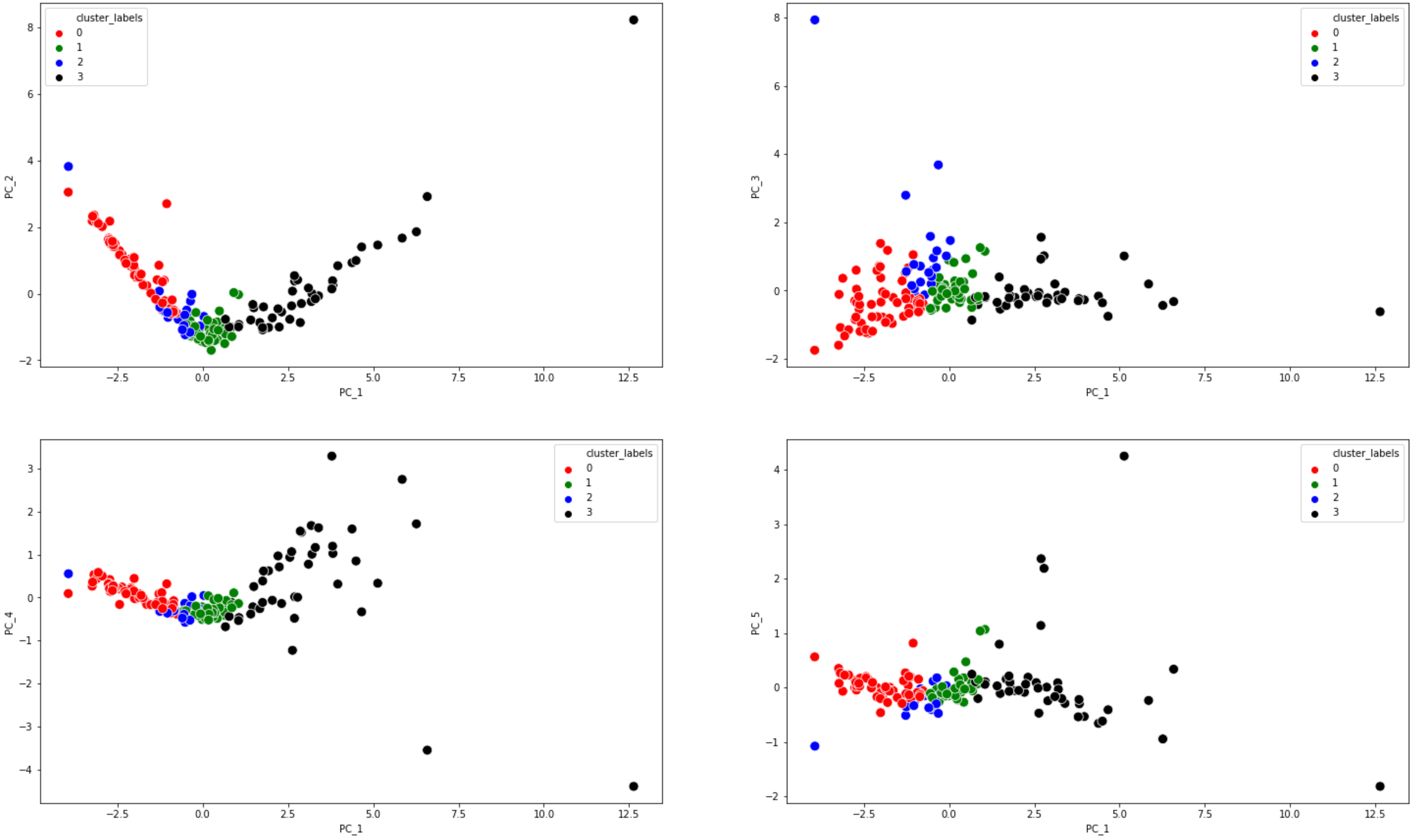
15. Hierarchical Clustering

- Dendogram based on Complete linkage:

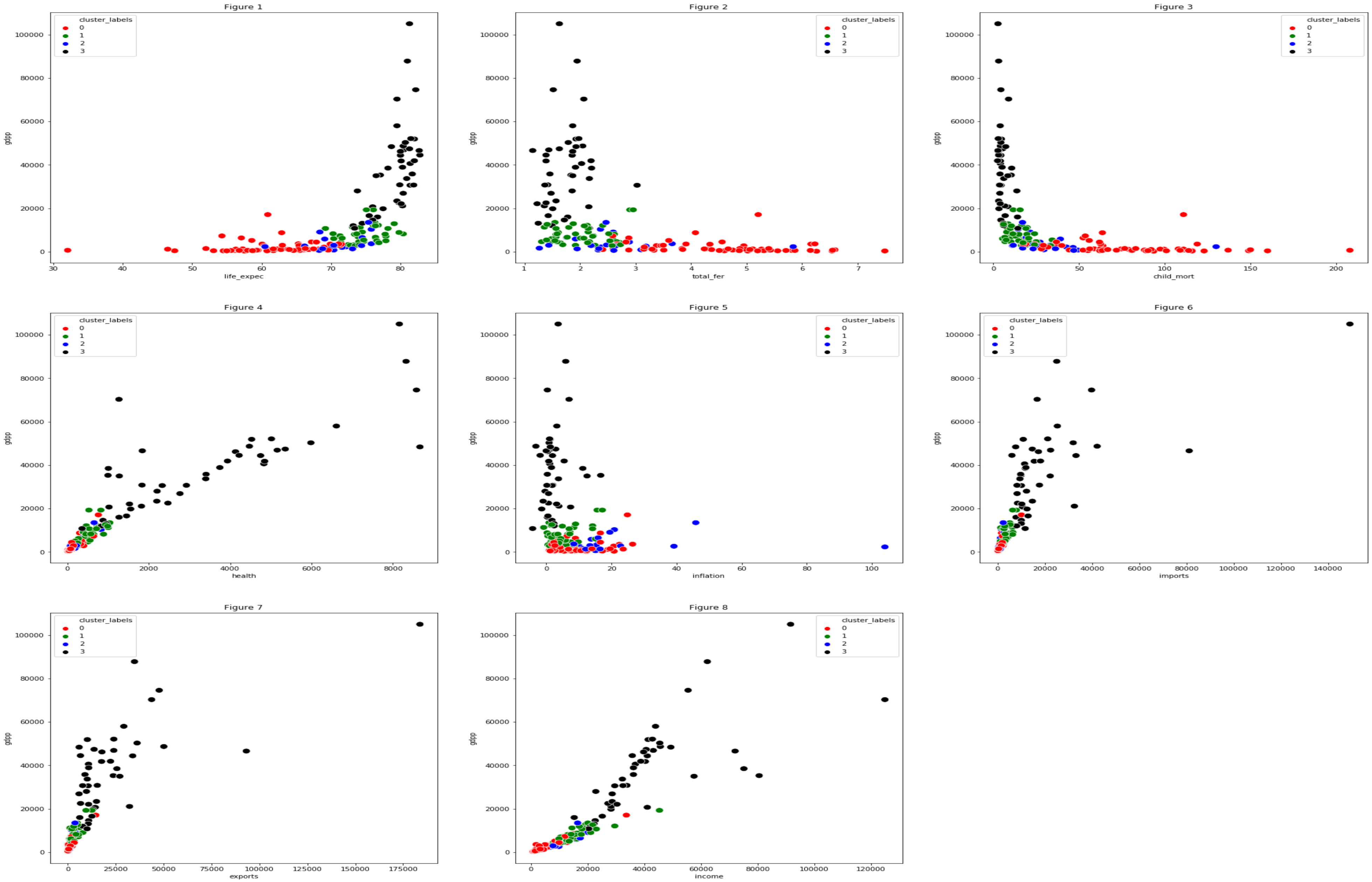


16. Based on the above chart, we can say that we have got 4 clusters, but this is different with what we get in K-Means clustering. Lets plot Scatter plot and Box plot to get more insight of the clusters formed.

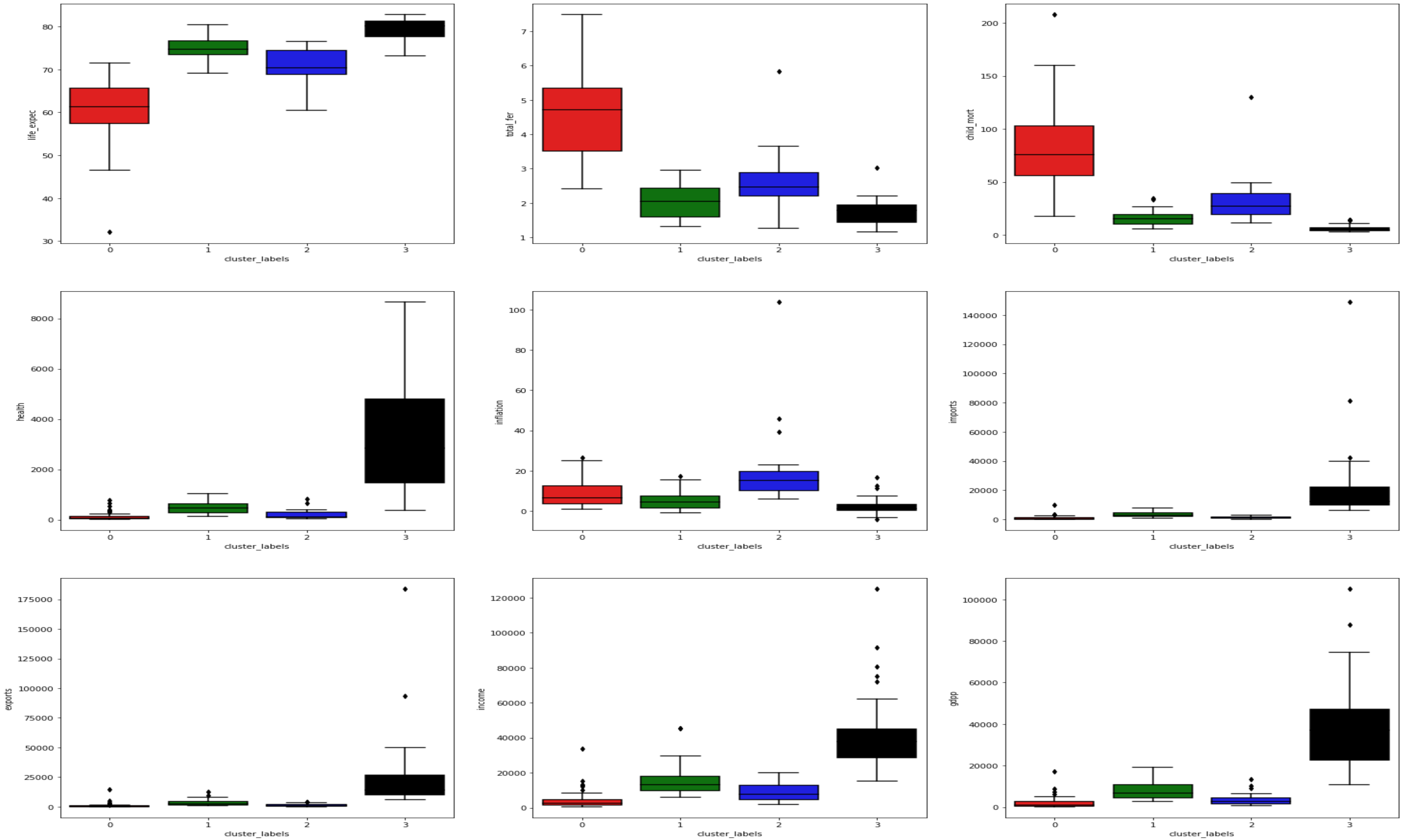
<u>Scatter Plot between PC_1-PC_2, PC_1-PC_3, PC_1-PC_4 and PC_1-PC_5</u>



17. Scatter plot on different variables taking hue as cluster_labels



18. Box plot on different variables taking hue as cluster_labels



Contd…

19. Based on the scatter plot and boxplot above, it has been clear that:
    - Cluster 0 represents under-developed countries
    - Cluster 1 represents emerging countries
    - Cluster 2 represents developing countries
    - Cluster 3 represents developed countries.
    - Cluster 0 has low life_expec, high_fertility, high child mortality, low health, low imports/exports, low income and low GDP. These all factors shows that Cluster 0 belongs to under-developed countries.
20. A new cluster has been added during running Hierarchical Clustering which shows much more insight of data. New cluster i.e. Cluster 1 is between developing countries and developed countries and are said to be emerging countries but we have to focus much on under-developed countries that need aid. After filtering out top 10 countries which are very much in need of aid are:
    - Benin, Burkina Faso, Central African Republic, Congo Dem. Rep., Guinea, Guinea-Bissau, Haiti, Mozambique, Niger and Sierra Leone

Result

- Hence it can be stated that both K Means and Hierarchical Clustering giving nearly same result if talking on under-developed countries. After filtering results in both types of clustering, we are getting same top 10 countries as listed below:
- Benin
- Burkina Faso
- Central African Republic
- Congo Dem. Rep.
- Guinea
- Guinea-Bissau
- Haiti
- Mozambique
- Niger
- Sierra Leone

These are the countries that are in the direst need of aid. We need to focus on these countries.